Change · Challenge · Opportunity
Injury & Disability Schemes Seminar

12 - 14 November 2017
Sofitel · Brisbane

Building better PPAC models

Prepared by Hugh Miller

Presented to the Actuaries Institute
Injury & Disability Schemes Seminar
12 – 14 November 2017
November

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9239 6100 **f** +61 (0) 2 9239 6170
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Building better PPAC models

Hugh Miller*

Abstract

While generalised linear models (GLMs) have achieved some level of acceptance in actuarial work, many reserving problems are still performed using traditional approaches in spreadsheets. A large part of this is that the perceived extra value from using a GLM is low, combined with the uncertainty of how a GLM compares to a traditional approach. This paper argues that the GLM approach is preferable, containing virtually all of the benefits of a traditional model and also permitting powerful extensions and improvements. These improvements to models using GLMs are largely unreported in the academic literature. The discussion is supported with theoretical and simulation results.

The paper focuses primarily on the PPAC model (payments per active claim), a very common valuation technique for long-tailed liabilities.

Keywords: Payments per active claim models; Generalised linear models; Active claims; Chain ladder; Reserving

1 Introduction

The last few decades has seen many changes in actuarial work due to the improvements in computational power and software. One area where this has met some resistance is the use of generalised linear models (or GLMs; see McCullagh & Nelder, 1989, and Dobson, 2002) to replace traditional models such as the the chain ladder for reserving triangles. Reserving triangles remain a key part of non-life insurance. There are a number of legitimate reasons for this resistance:

- Traditional average methods can be easily performed in spreadsheet software such as Microsoft Excel. This means the dataset, assumptions and forecast are easy to inspect simultaneously; changes to assumptions can be easily made and their impacts quickly seen; the model is interpretable to other users;
- Traditional average methods are seen to be relatively fast to implement.
- The "value-add" from using a generalised linear model is often unclear, particularly to non-technical users of reserving estimates.
- There are some practical issues with using GLMs, including treatment of zeros in the data and the ability to impose prior knowledge on model parameters.
- A familiarity bias, where conversion to a GLM can appear daunting to actuaries not experienced with linear models.

*Correspondence to: Hugh Miller, Taylor Fry, Level 22, 45 Clarence St, Sydney, 2000, Australia. Email: hugh.miller@taylorfry.com.au

This paper discusses the use of GLMs for these types of reserving models. It primarily focuses the third and fourth items listed above; GLMs can do everything a traditional model can do, plus more, and many of the practical issues can be overcome. However the other items are also addressed, arguing that properly implemented GLMs can be easy to use and standard fits can be implemented quickly.

To make the discussion concrete, a specific type of model, the "payment per active claim" or PPAC model will be discussed. This has been chosen due to its common use in modelling long tailed benefits such as workers' compensation, a difficult class for reserving. Most of the comments are applicable to other traditional actuarial reserving models.

This paper is primarily pedagogical; it seeks to be practical in applying GLMs to reserving problems, bridging actuarial theory and practice. However, there are some novelties. Most of the ideas in Section 4 are new to the actuarial literature. Further, some claims as to best practice are supported by theoretical results in the technical appendix.

The remainder of the paper is structured as follows. Section 2 introduces PPAC models, GLMs and example datasets in some detail. Section 3 discusses some of the considerations for all types of PPAC models. Section 4 covers a number of additional considerations for when PPAC models are built using GLMs. Section 5 provides some further discussion and conclusions. Finally, the theoretical appendix provides some statistical results that underlie some of the discussion throughout the paper.

2 Background

2.1 Data We assume a triangle of historical data, with rows representing accident years $i = 1, \dots, I$, columns development years $j = 1, \dots, J$ and payment year $k = i + j = 1, \dots, K$. The general aim of loss reserving aim is to "complete the triangle", to determine payments that will occur in the future related to accidents incurred in the past. For each cell in the triangle we have two pieces of information; the number of people "active" on that benefit, n_{ij}, and the amount of payments, Q_{ij}. Payments are often inflation-adjusted at the beginning of the analysis. While the definition of active can vary depending on context, one simple and oft-used definition is the presence of any payment in that year.

The chosen interval above is years, but this is of course arbitrary and different time increments can be adopted. Other possible extensions to the data, not pursued in this paper, include:

- The inclusion of other *claimant* variables. One obvious example is claimant age, as different ages have different rates of continuance on benefit as well as permitting explicit retirement or death decrements if desired. If age were added, then each cell of data (i, j) is split into multiple observations; one for each observed age group in the cell.
- The inclusion of other *time related* variables. The behaviour of beneficiaries is often correlated to macro-economic variables such as unemployment rates

or growth in GDP. These are relatively straightforward to "attach" these modelling variables to the appropriate diagonals.

2.2 PPAC models The payment per active claim is defined as total payments in the cell divided by the number of active claims

$$P_{ij} = Q_{ij}/n_{ij} \tag{2.1}$$

A PPAC based valuation is then made up of two models, described below.

The *continuance rate model* relates to the rate at which active claims become inactive or closed. It involves the estimation of continuance rate factors c_j, which are typically assumed to vary by development period but not by accident year.

$$n_{ij} \approx c_j n_{i,j-1} \tag{2.2}$$

The c_j are estimated based on historical ratios of $n_{ij}/n_{i,j-1}$ across the various accident years i. In practice the observed continuance rates will not always be constant across accidents years, either due to random fluctuations or systematic evolution over time. For this reason we express equation (2.2) as approximate. Noise is generally dealt with via averaging continuance rates over a number of accident periods, while systematic evolution requires some assessment of the trend before selecting c_j for future periods.

The *payment model* relates to the average payment level per active claim. Again, these factors P_j is typically assumed to vary by development period but not by accident year.

$$P_j \approx P_{ij} = \frac{Q_{ij}}{n_{ij}} \tag{2.3}$$

Observed payment levels will not necessarily be constant down a column of the triangle due to both noise and systematic effects. The most common systematic effect is superimposed inflation (SI), where succcessive payment years (diagonals) tend to see increases.

PPAC models have proven to be a popular form of reserving model, allowing careful study of numbers of active claims, and are useful where there are consistently sized payments made over an extended period of time. They are used in many countries, but are particularly popular in Australia to value long term injury claims such as workers' compensation.

PPACs are but one example of reserving triangles, which have been studied extensively in the actuarial literature. Taylor (1986) gives an overview of early work on reserving triangles. The Mack (1993) model established a distribution-free formula for the standard error of chain ladder reserves. England & Verall (2002) extended this to a fuller framework of stochastic claim reserving. The textbook by Taylor (2000) gives a relatively modern theoretical overview to loss reserving.

Accident Year	Development year 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1995	105	73	55	47	38	35	33	31	30	29	29	29	29	29	28	27	26	22	22
1996	110	77	56	43	38	33	31	31	31	31	29	29	27	27	25	25	24	24	23.3
1997	115	85	63	55	50	42	41	40	40	39	37	37	35	35	35	33	32	31.0	30.1
1998	120	85	67	52	41	34	31	29	28	26	26	26	25	25	25	23	22.3	21.6	21.0
1999	126	87	68	50	42	40	38	33	32	31	30	30	29	29	28	27.2	26.3	25.6	24.8
2000	132	85	63	52	48	40	39	37	36	35	35	34	33	31	30.1	29.2	28.3	27.4	26.6
2001	138	94	78	59	48	45	45	43	40	37	36	36	36	34.9	33.9	32.9	31.9	30.9	30.0
2002	144	97	72	60	52	44	43	41	40	40	39	39	38.1	36.9	35.8	34.7	33.7	32.7	31.7
2003	151	112	80	61	54	50	49	46	45	44	42	41.7	40.7	39.5	38.3	37.2	36.0	35.0	33.9
2004	158	114	91	77	63	55	52	49	46	46	44.8	44.5	43.4	42.1	40.9	39.6	38.5	37.3	36.2
2005	165	112	89	75	62	54	51	48	47	45.9	44.7	44.4	43.3	42.0	40.8	39.6	38.4	37.2	36.1
2006	173	113	87	72	58	49	46	43	41.6	40.6	39.6	39.3	38.4	37.2	36.1	35.0	34.0	32.9	32.0
2007	181	129	105	88	74	66	63	59.2	57.3	55.9	54.5	54.1	52.8	51.2	49.7	48.2	46.7	45.3	44.0
2008	190	131	103	75	63	57	53.9	50.7	49.0	47.9	46.7	46.3	45.2	43.9	42.5	41.3	40.0	38.8	37.7
2009	199	137	106	79	66	58.0	54.9	51.6	49.9	48.8	47.5	47.2	46.0	44.7	43.3	42.0	40.8	39.5	38.3
2010	208	138	104	90	74.8	65.8	62.3	58.5	56.6	55.3	53.8	53.5	52.2	50.6	49.1	47.6	46.2	44.8	43.5
2011	218	144	104	82.6	68.7	60.4	57.1	53.7	51.9	50.7	49.4	49.1	47.9	46.5	45.1	43.7	42.4	41.1	39.9
2012	228	156	118.3	93.9	78.1	68.7	65.0	61.0	59.1	57.7	56.2	55.8	54.5	52.8	51.2	49.7	48.2	46.8	45.4
2013	239	161.1	122.1	97.0	80.6	70.9	67.1	63.0	61.0	59.6	58.0	57.6	56.3	54.6	52.9	51.3	49.8	48.3	46.9
Continuance rate calculations																			
All		0.6882	0.7672	0.8042	0.8434	0.8810	0.9574	0.9439	0.9696	0.9728	0.9712	0.9962	0.9683	0.9888	0.9724	0.9558	0.9647	0.9200	1.0000
Last 4		0.6741	0.7582	0.7943	0.8312	0.8794	0.9464	0.9394	0.9674	0.9766	0.9744	0.9929	0.9762	0.9836	0.9741	0.9558	0.9647	0.9200	1.0000
Last 2		0.6726	0.7376	0.8048	0.8377	0.8978	0.9478	0.9381	0.9588	0.9890	0.9643	1.0000	0.9857	0.9677	0.9815	0.9333	0.9655	0.9200	1.0000
Selected		0.6741	0.7582	0.7943	0.8312	0.8794	0.9464	0.9394	0.9674	0.9766	0.9744	0.9929	0.9762	0.9700	0.9700	0.9700	0.9700	0.9700	0.9700

Table 1: Chain ladder continuance rate model, synthetic data

Accident Year	Development year 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1995	1,000	2,275	2,287	2,752	2,264	2,460	2,627	3,350	2,780	4,653	4,333	3,220	3,569	3,966	4,244	3,995	4,950	5,231	4,075
1996	1,131	2,190	2,269	2,102	2,647	2,464	2,489	2,878	3,318	4,139	2,807	4,083	3,665	4,602	4,417	4,313	5,132	4,774	4,075
1997	1,189	2,122	2,335	2,094	3,370	2,808	2,880	3,008	3,266	3,805	3,961	3,567	4,417	3,924	4,520	4,772	4,441	4,993	4,075
1998	1,166	2,511	2,284	2,244	2,550	2,942	2,878	3,157	3,070	3,759	3,835	4,554	3,511	4,612	4,726	4,810	4,805	4,993	4,075
1999	687	2,091	2,726	2,599	3,853	2,973	3,223	4,261	3,379	3,983	3,485	4,059	4,246	4,851	4,086	4,480	4,805	4,993	4,075
2000	940	2,381	2,718	2,389	3,425	3,110	3,499	3,332	3,290	3,890	4,274	4,738	3,983	4,678	4,435	4,480	4,805	4,993	4,075
2001	1,355	2,732	2,758	2,768	3,204	3,207	3,492	3,401	3,964	4,004	4,499	4,266	3,904	4,486	4,435	4,480	4,805	4,993	4,075
2002	1,195	2,453	3,230	3,344	3,077	3,291	3,539	3,865	4,379	4,032	4,120	4,472	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2003	922	2,509	2,600	3,146	3,428	3,215	3,447	4,529	4,167	4,215	4,695	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2004	1,307	2,577	2,965	3,201	3,147	3,760	3,564	3,665	3,685	4,706	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2005	1,528	2,899	3,055	3,170	3,147	4,245	4,062	3,435	4,198	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2006	1,236	2,823	3,193	3,735	3,735	3,894	3,745	4,171	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2007	1,517	2,862	3,105	3,227	3,249	4,132	4,088	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2008	1,666	3,152	3,041	3,386	3,392	3,911	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2009	1,632	2,773	3,524	3,754	4,041	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2010	1,599	3,148	3,146	3,520	3,592	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2011	1,658	3,165	3,321	3,468	3,592	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2012	1,658	3,954	3,260	3,468	3,592	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
2013	1,631	3,281	3,260	3,468	3,592	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
Payment per active claim, no superimposed inflation assumption. Averages are volume weighted																			
All	1,376	2,791	2,926	3,059	3,282	3,418	3,435	3,621	3,656	4,135	4,051	4,127	3,925	4,418	4,397	4,480	4,805	4,993	4,075
Last 4	1,637	3,281	3,260	3,468	3,592	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075
Last 2	1,645	3,575	3,233	3,629	3,724	4,030	3,943	3,783	3,944	4,466	4,418	4,373	3,942	4,762	4,388	4,787	4,737	4,993	4,075
Selected		3,281	3,260	3,468	3,592	4,052	3,879	3,936	4,098	4,260	4,404	4,394	3,926	4,486	4,435	4,480	4,805	4,993	4,075

Table 2: Payments model, synthetic data

2.3 Traditional spreadsheet PPAC implementations Much of the current practice of these types of models still use chain ladder based approaches. Tables 1 and 2 show an example of this on a synthetic dataset. We carry this example through the paper for concreteness. In each table the upper triangle is historical, while the lower half are projections based on the selections for continuance rates (c_j) and payment level (P_j). The projected numbers of actives and payment levels can then be multiplied to obtain all projected future cash flows. We make a few additional comments about the example:

- The "All" continuance rate estimates are the overall (volume-weighted) average continuances for all historical entries in the column. If $\mathcal{H}$ denotes the set of i, j that are historical ($i + j$ are equal to or less than the current payment year) then the estimate is:

$$\hat{c}_j = \frac{\sum_{i:(i,j)\in\mathcal{H}} n_{ij}}{\sum_{i:(i,j)\in\mathcal{H}} n_{i,j-1}}\,, \quad \text{for } j \geq 1. \tag{2.4}$$

  The "Last 4" estimate for c_j is the equivalent sum, but only using the four most recent diagonals, so for instance the first number in the row is calculated as

$$\frac{137 + 138 + 144 + 156}{199 + 208 + 218 + 228} = 0.6741\,.$$

  Similarly "Last 2" uses the last two diagonals. The use of "Last X" averages is a very common way to have the adopted factors emphasise the most recent trends. The projected values use the last 4 average for the projection for development periods 1 to 12.

- The 0.97 selected for development periods 13 to 18. It is common to "pool" averages across development periods, particularly in averages of higher uncertainty (continuance rates close to 1, or low numbers of observed claims).

- While not done here, it is common practice to "smooth" the leading diagonal (with reference to prior observations in the row). This reflects a belief that unusually high (or low) continuance in a cell is often followed by reversion via lower (or higher) continuance thereafter.

- The projection of actives starts from the leading diagonal and applies the selected continuance factor assumptions according to Equation 2.2.

- The "All" payment level rates use a (volume-weighted) average:

$$\hat{P}_j = \frac{\sum_{i:(i,j)\in\mathcal{H}} P_{ij} n_{ij}}{\sum_{i:(i,j)\in\mathcal{H}} n_{ij}}\,. \tag{2.5}$$

  The "last X" rows are the equivalent averages using only the last few rows of data. Again, this is common practice to emphasise recent trends.

- There appears to be superimposed inflation present in the payments triangle, with earlier rows appearing lower than cells below. The selection shown in Table 2 does not allow for this feature, but we discuss superimposed inflation in Section 4.2.

- There are a significant number of active claims at development year 18. In practice the projection would be extended to some larger number of development periods to estimate payments in the tail of the liability. Continuance and payment level assumptions would generally be extrapolated from the latest observed development periods.

2.4 GLMs for PPAC models A GLM for continuance rate modelling is a typically a model of the form:

$$c_{ij} = g^{-1}(X_{ij}^T \beta) + error\,, \tag{2.6}$$

where

X_{ij}	=	p-vector of predictors (or covariates) corresponding to cell (i, j)
g^{-1}	=	the inverse of the link function. The log-link $g^{-1}() = exp()$, is common to ensure positive continuance rates
$error$	=	An error term with zero mean

The response of a GLM (and thus the distribution of the error term) is assumed to be taken from the exponential family of distributions (see Fahrmeier & Tutz, 1994 or Clark & Thayer, 2004), which will then determine the distribution for the error term. The β are then estimated by maximum likelihood. It is common for the first element of X_{ij} to equal 1 for all i, j combinations, corresponding to an intercept term. A input dataset for a GLM has one row per "observation", so the data triangle must be converted into a flat file - see Table 3 for an example layout of the synthetic continuance rate triangle with illustrative covariates. A weight term for each observation w_{ij} can be included in the maximum likelihood calculation to reflect the relative importance of each observation. For continuance rate models the number of active claims in the previous year, $w_{ij} = n_{i,j-1}$, forms a natural weight. This choice of weight produces volume-weighted estimates of the type seen in Section 2.3

Of particular interest to this paper is the choice of covariates for the model. Let $\mathbb{I}(\mathcal{A})$ be the indicator function that takes the value 1 if $\mathcal{A}$ is true and zero otherwise. If :

$$X_{ij} = \{\mathbb{I}(j = 1), \mathbb{I}(j = 2), \ldots, \mathbb{I}(j = J)\}\,,$$

then one parameter estimate will be made for each development period. This setup has the useful property of recovering the average chain ladder continuance factors calculated in Table 1 (see also Result 1 in the technical appendix for the proof).

The GLM for payment levels is very similar in application:

$$P_{ij} = g^{-1}(X_{ij}^T \beta) + error\,, \tag{2.7}$$

A good default weight is $w_{ij} = n_{ij}$, the number of active claims.

Index		Weight	Response	Covariates			
AY	**DY**	**Prev_active**	**Continaunce**	**Intercept**	**I(j=1)**	**I(j=2)**	**...**
1995	1	105	0.695	1	1	0	
1995	2	73	0.753	1	0	1	
1995	3	55	0.855	1	0	0	
1995	4	47	0.809	1	0	0	
1995	5	38	0.921	1	0	0	
1995	6	35	0.943	1	0	0	
1995	7	33	0.939	1	0	0	
1995	8	31	0.968	1	0	0	
1995	9	30	0.967	1	0	0	
1995	10	29	1.000	1	0	0	
1995	11	29	1.000	1	0	0	
1995	12	29	1.000	1	0	0	
1995	13	29	1.000	1	0	0	
1995	14	29	0.966	1	0	0	
1995	15	28	0.964	1	0	0	
1995	16	27	0.963	1	0	0	
1995	17	26	0.846	1	0	0	
1995	18	22	1.000	1	0	0	
1996	1	110	0.700	1	1	0	
1996	2	77	0.727	1	0	1	
1996	3	56	0.768	1	0	0	
1996	4	43	0.884	1	0	0	
1996	5	38	0.868	1	0	0	
...		...			...		

Table 3: Flat file format suitable for GLMs, for the synthetic continuance rate model example

2.5 GLM equivalents for traditional PPAC formulations One important and straightforward result is that the standard types of selections shown in Tables 1 and 2 are easy to reproduce using a GLM. The previous section introduced a simple development year parameterisation:

$$\textbf{Model 1:} \quad X_{ij} = \{\mathbb{I}(j=1), \mathbb{I}(j=2), \ldots, \mathbb{I}(j=J)\}\,. \qquad (2.8)$$

Most statistical packages allow this model to be expressed in a shorthand fashion, with the covariate input a single categorical development year term. The package will convert this into the required indicators and report back a parameter estimate for each level. For instance in SAS the required model statement is simply `Continuance = DY`, plus a statement to identify `DY` as categorical.

A "Last X" type estimate can be obtained by interacting the development year effect with a payment year effect. If $k = K$ is the latest payment year, then the last 4 estimates can be found using a payment year indicators that pick up the last four diagonals and everything else:

Model 2:

$$\begin{aligned} X_{ij} \;=\; & \{\mathbb{I}(j=1)\mathbb{I}(k \geq K-3), \mathbb{I}(j=2)\mathbb{I}(k \geq K-3), \ldots, \mathbb{I}(j=J)\mathbb{I}(k \geq K-3), \\ & \mathbb{I}(j=1)\mathbb{I}(k < K-3), \mathbb{I}(j=2)\mathbb{I}(k < K-3), \ldots, \mathbb{I}(j=J)\mathbb{I}(k < K-3)\} \end{aligned}$$

Note this formulation has some redundancy in the sense that the user is not generally interested in the parameters corresponding to the older payment years[1]. These "bystander" terms are required to ensure the last diagonals are not impacted by the older years in the maximum likelihood calculation. Setting the weights to these older years (or deleting them from the flat modelling file altogether) would produce the last 4 averages without the need for these bystander terms. The SAS model statement to produce this is `Continuance = DY*PY_pre_last4 DY*PY_last4`, with the PY terms the required payment year indicator functions in Model 2 above.

It is also possible to calculate continuance rate averages across development periods. In our synthetic example, Table 1 adopts a single estimate for DY 13 and beyond. This can be done in a GLM setup too:

Model 3:

$$\begin{aligned} X_{ij} &= \{\mathbb{I}(j=1)\mathbb{I}(k \geq K-3), \ldots, \mathbb{I}(j=12)\mathbb{I}(k \geq K-3), \mathbb{I}(j \geq 13)\mathbb{I}(k \geq K-3), \\ &\quad \mathbb{I}(j=1)\mathbb{I}(k < K-3), \ldots, \mathbb{I}(j=12)\mathbb{I}(k < K-3), \mathbb{I}(j \geq 13)\mathbb{I}(k < K-3)\} \end{aligned}$$

DY	Continuance rate models						Payment level models			
	Model 1 - All		Model 2 - Last 4		Model 3		Model 1 - All		Model 2 - Last 4	
	β	Exp(β)	Beta	Exp(Beta)	Beta	Exp(Beta)	Beta	Exp(Beta)	Beta	Exp(Beta)
0							7.2269	1,376	7.4007	1,637
1	-0.374	0.688	-0.394	0.674	-0.394	0.674	7.934	2,791	8.096	3,281
2	-0.265	0.767	-0.277	0.758	-0.277	0.758	7.981	2,926	8.089	3,260
3	-0.218	0.804	-0.230	0.794	-0.230	0.794	8.026	3,059	8.151	3,468
4	-0.170	0.843	-0.185	0.831	-0.185	0.831	8.096	3,282	8.186	3,592
5	-0.127	0.881	-0.129	0.879	-0.129	0.879	8.137	3,418	8.307	4,052
6	-0.044	0.957	-0.055	0.946	-0.055	0.946	8.142	3,435	8.263	3,878
7	-0.058	0.944	-0.063	0.939	-0.063	0.939	8.195	3,621	8.278	3,936
8	-0.031	0.970	-0.033	0.967	-0.033	0.967	8.204	3,657	8.318	4,098
9	-0.028	0.973	-0.024	0.977	-0.024	0.977	8.327	4,135	8.357	4,259
10	-0.029	0.971	-0.026	0.974	-0.026	0.974	8.307	4,051	8.390	4,404
11	-0.004	0.996	-0.007	0.993	-0.007	0.993	8.325	4,127	8.388	4,394
12	-0.032	0.968	-0.024	0.976	-0.024	0.976	8.275	3,925	8.275	3,926
13	-0.011	0.989	-0.017	0.984	-0.034	0.967	8.394	4,418	8.409	4,486
14	-0.028	0.972	-0.026	0.974	-0.034	0.967	8.389	4,398	8.397	4,435
15	-0.045	0.956	-0.045	0.956	-0.034	0.967	8.407	4,480	8.407	4,480
16	-0.036	0.965	-0.036	0.965	-0.034	0.967	8.477	4,805	8.477	4,805
17	-0.083	0.920	-0.083	0.920	-0.034	0.967	8.516	4,993	8.516	4,993
18	0.000	1.000	0.000	1.000	-0.034	0.967	8.313	4,075	8.313	4,075

Table 4: GLM fit results for models described in Section 2.3. Estimates are equivalent to those in Tables 1 and 2.

This discussion leads naturally to the first (well-known) result that traditional "All" and "Last X" estimates for continuance rate and payment level models can be exactly reproduced using GLMs. Formally, this is a consequence of Result 1, in the technical appendix.

[1]We note however that while not needed for the projection, fitting to the older part of the triangle can have uses. It gives insight into past changes in the scheme, and allows a better understanding of variability and distribution of errors.

The table also shows the payment level model results for the Model 1 and 2 formulations. Again, these match the traditional chain ladder estimates. We make a few further comments on the GLM equivalences:

- The parameter estimates for formulations that have only indicator functions are independent of distribution choice; the GLM will always return volume weighted averages. Parameter estimates will vary by distribution choice in the presence of more complex effects such as linear trends. The distribution choice also has significant implications for the standard errors and statistical diagnostics related to the GLM fit.

- The results shown for continuance rates are a Poisson model with log-link, the response c_{ij} and weights $n_{i,j-1}$. It turns out there is an equivalent formulation for Poisson log-link models using n_{ij} as the response and $\log(n_{i,j-1})$ as an offset. This alternative formulation can occasionally be more convenient. Offsets are discussed further in Section 4.6 in the context of incorporating continuance rate information into the payment model.

3 Considerations for all types of PPAC models

3.1 Net continuance versus deactivation and re-activation models One common feature of the claims analysis is that claims under study will deactivate and then reactivate. If actives are defined by the existence of a payment in the year, this pattern may correspond to a claim formally closing and reopening, or a period where the claim remains open but no payments were made. The presence of reactivations raises the question of whether it is better to model:

- A "net continuance rate", which calculates the rate of continuance allowing simultaneously for activations and deactivations

- Separate continuance rate type models for deactivations and reactivations, which can be combined to estimate total future actives

There are some natural attractions to having separate models. The extra detail allows a better understanding of the how claims evolve over time, and so emerging trends in one of these rates can be easier to spot. Further, even if the rates of deactivation and reactivation are stable over time, then the pattern in the net deactivation has to implicitly allow for the relative change in numbers active and inactive, making the net continuance model more challenging.

However the advantages of separate modelling are usually dwarfed by a key disadvantage - instability in projection. Unless the activation and deactivation rates are perfectly matched, then any long term projection (particularly those beyond the observed development length) will often extrapolate poorly. For instance, when the difference between the deactivation rate and reactivation is set slightly too low, an unreasonably large liability can occur. We explore the theoretical reasons for this behaviour in the technical appendix; under mild assumptions the variability of the joint model will always be higher, which we've also verified via simulation.

For this reason we generally prefer the modelling of net continuance rate, as it gives more direct control over how actives are extrapolated. The comment applies equally to chain ladder and GLM approaches. We'll assume a net continuance rate model for most of the discussion below.

3.2 Dealing with zeroes in the actives triangle Often zero cells occur in a continuance rate triangle, meaning that a particular accident-development period combination had no active claims. This is particularly common for triangles where there is an extensive history, or when the time unit is small (such as a monthly continuance triangle). This can cause issues in continuance rate estimation, both in traditional chain ladder approaches (a zero in the denominator can cause errors), and GLMs (the continuance rate in the cell is undefined).

For both approaches the issue can be ignored if the subsequent development period is also zero. In this case the observation can be ignored for the purpose of estimation, and continuance rates estimated using other observations. However, in cases where there is a non-zero entry (as can happen when modelling net continuance rate), the implied continuance rate is infinite and unless the observation is altered or deleted the GLM will typically fail to estimate. For traditional chain ladder models, this is usually handled by calculating an average continuance rate over enough cells such that the denominator is non-zero. This gives an unbiased continuance rate estimate across those cells.

For GLMs, a convenient trick is to set all zeros to some small fraction (such as 0.001). This leads to a large continuance rate for the subsequent cell, counterbalanced by a low weight assigned to it (when using $n_{i,j-1}$ as the weight). It can be shown that as this fraction approaches zero the GLM parameter estimates converge and are asymptotically unbiased; see Result 3 of the technical appendix. One limitation of this approach relates to the estimation of the scale parameter. This is the global estimate of variance for the model (see McCullagh & Nelder, 1989) which is usually estimated in one of two ways. The Deviance estimate is stable when the zeros are replaced by small numbers, but the Pearson estimate of scale is unstable.

4 Considerations for GLM-based PPAC models

4.1 Alternative formulations The setup of Section 2.4 used the continuance rate as the response variable. A common alternative is to model the number of active claim numbers directly, using a cross-classified structure:

$$n_{ij} = g^{-1}(X_{ij}^T \beta) + error, \tag{4.1}$$

For example, a common theoretical structure is to allow X_{ij} to have accident and development year main effects:

Model 4:

$$\begin{aligned} X_{ij} &= \{\mathbb{I}(j=1), \mathbb{I}(j=2), \ldots, \mathbb{I}(j=J), \\ &\quad \mathbb{I}(i=1), \mathbb{I}(i=2), \ldots, \mathbb{I}(i=I)\} \end{aligned}$$

With a log link this is often characterised as

$$X_{ij} \approx \alpha(i)\beta(j).$$

See for instance Chapter 7 of Taylor (2000).

This formulation is not, in general, equivalent to the continuance rate model of (2.8). If instead the response was the *change* in actives $n_{ij} - n_{i,j-1}$, then an equivalence would in fact occur - see for instance England and Verrall (2002). In most cases the results of (4.1) will be fairly similar to (2.8) - differences depend on the extent to which there are unusually high or low entries on the latest diagonal. This alternative formulation can be useful however, as it permits more direct control over the *level* of active claims, rather than its decay rate. For example, it is slightly easier to impose constant active numbers over regions of the triangle using this setup.

4.2 Superimposed inflation Superimposed inflation (SI) relates to the rate of increase in claims cost by payment period $k = i + j$. In traditional spreadsheet based approaches estimation of SI can be a challenge, because there is rarely a formal basis for what a "best" estimate of SI should look like. In a GLM context, SI is easy to estimate by adding a continuous payment year term to the model specification. This approach will then simultaneously estimate the development factors and the observed SI. So for instance a "last 4" type payment model would take the form below.

Model 5:

$$\begin{aligned} X_{ij} &= \{k, \mathbb{I}(j=1)\mathbb{I}(k \geq K-3), \mathbb{I}(j=2)\mathbb{I}(k \geq K-3), \ldots, \mathbb{I}(j=J)\mathbb{I}(k \geq K-3), \\ &\quad \mathbb{I}(j=1)\mathbb{I}(k < K-3), \mathbb{I}(j=2)\mathbb{I}(k < K-3), \ldots, \mathbb{I}(j=J)\mathbb{I}(k < K-3)\} \end{aligned}$$

We make some further comments regarding the estimation of SI:

- Assuming the stochastic setup of the GLM is fair (in particular correct choice of distribution), then the standard error of the SI parameter is useful to indicate the uncertainty of the SI.
- A relatively straightforward way to visually inspect for SI is to conditionally colour cells of the triangle relative to the average of that development column. Patterns of colours may motivate separating SI assumptions in different regions.
- Different payment year linear splines can be tested to estimate SI over different time periods. This could either test for the presence of any SI over certain time periods, or test for any change in the level over time.

Estimation of superimposed inflation in a GLM context is thus relatively straightforward. Analysis of past superimposed inflation still requires some belief on whether trends will continue; this is typically only gained through deeper understanding of an insurance scheme and drivers of cost. It is also possible to go further with superimposed inflation estimation:

- Changes in superimposed inflation over time can be tested by adding linear splines of payment period to the model. These tests can be for a departure from the previous rate of superimposed inflation, or tests against the hypothesis that the rate has fallen to zero.

- If a particular rate of superimposed inflation is desired, this can be imposed via an offset (similar to Section 4.6). For log-link models, this offset would take the form of $k \log(1 + s)$, with s the desired rate of SI.

4.3 Setup, fast feedback and diagnostics Traditional GLM diagnostics include:

- Overall measures of fit, such as the AIC or BIC

- Observation level measures of fit, such as the various types of residuals

- Parameter level statistics, such as the estimate, standard error and related p-value

- Diagnostic plots, such as relativity plots by predictor or actual versus expected averages by predictor

However these tend not to be the most useful suite of diagnostics in a reserving context. Ultimately this is because a reserving problem is "extrapolation" (applying estimates to combinations of (i, j) outside the range of experience) rather than "interpolation" (estimating cases that fall within the range of previously observed cases). This issue has a few consequences. First, overall measures of fit encourage improving the fit on the historical data, whereas the analyst is primarily interested in the projection and the fit on the last few diagonals. For example, in the "Last 4" setup of model 2, a large number of extra terms could be added to the $\mathbb{I}(k < K - 3)$ component of the model, improving the fit (and diagnostics) on the older diagonals. However none of these effects would have any impact on the fit on the last 4 diagonals, and no impact on the projection either. Thus the overall measure of fit is not always a useful diagnostic.

Second, focusing observation level diagnostics such as residuals does not necessarily improve the model. As an extreme example, if a particular cell has a very large residual, an analyst might be tempted to add an parameter that affects only that cell. The residual would then be improved, but the information from that cell no longer contributes to the extrapolation of the triangle. More generally, only a subset of the parameters added to the model influence the future extrapolation region, and some of these terms will actually "remove" information feeding the projection, which tends to increase the variability.

Third, standard parameter level significance tests are not targeted to the impact on the extrapolation. This is particularly important for the continuance rate model, as an adopted continuance factor c_j will affect all development periods from j onwards. It is thus more useful to use a measure of "significance" in terms of its impact on the projection, rather than the standard view within the historical development column.

Given the above comments, what should an analyst be using to assess a fit? We are not sure that there is a best practice developed, but the following outputs should be considered:

- **Direct application of a fit to the projection half of the triangle:** It is usually not hard to automate the GLM so that a matrix of all historical and projected numbers are produced with every fit. This is an important element, as it immediately provides feedback on the reasonableness of the model and aids in some of the other comparisons below.

- **Direct comparison of projection to a reference model:** A reference model might be the one used in a previous valuation, or one produced using a "default" structure such as the ones described in this paper. Producing a projection triangle of ratios (current model divided by reference) will highlight which areas of the fit is extrapolating differently, and allows an analyst to quickly identify areas to check. These comparisons can be made at a cell level, as well as an accident year and aggregate level.

- **Monitoring of historical and projected cumulants:** While it is important to look at series of individual columns, it is also useful to examine how the columns accumulate. This is particularly useful in a continuance rate model; even if for each given j the adopted c_j might look plausible compared to the historical c_{ij}, the products $c_j c_{j+1} \cdots c_{j+t}$ for various j and t might show more misfit compared to their historical equivalents. One motivation for this type of checking is because successive continuance rates in historical data are often not independent, so naive "last X" fits can fail to recognise that the observed values on the leading diagonal might depend on the prior experience of that accident period.

  Similarly for the payment model, looking at sums over bands of development periods can reveal systematic biases in estimation.

- **Comparison of residuals across continuance and payment models:** It is not unusual to see correlations between residuals in equivalent cells in the continuance and payment triangles. Part, but probably not all, of this is a natural consequence of the PPAC model definition; see Section 4.6. Monitoring the presence of correlations reduces the chance of inconsistently applied assumptions. For example, if an unusually low continuance rate is viewed as an outlier and ignored, retaining the an unusually low payment level in the corresponding cell is potentially inconsistent.

Applying these tools, as well as correct use of the usual GLM diagnostics, should allow a fitting process more tightly integrated with reserve estimation.

4.4 Distribution Choice and stochastic interpretations To this point we have not considered the distributional considerations and stochastic interpretation of the GLM approach. However, to many users this represents a core feature of the GLM approach. Assuming the distributional assumptions are correctly set:

- The statistical significance of parameters can be estimated, and it is easy to test whether two continuance factors are likely to be different or not.
- The uncertainty of specific estimates can be calculated and reported
- The uncertainty of the total outstanding claims can be calculated, either theoretically (see England and Verrall, 2002) or via the bootstrap (Taylor, 2000, or England and Verrall, 1999).

Generally an over-dispersed Poisson distribution or a negative binomial distribution is appropriate for a continuance rate model. The range of possible distirbution for the payment model is wider, with gamma and Tweedie models both common but not exhaustive. However, in both cases some care is needed in checking for heteroscedasticity - often different portions of the triangle will show different levels of variability. This is particularly true for different development periods and has a number of causes:

- The changing relative influence of deactivations and (re)activations, in the case where a net continuance rate model has been fit.
- Some claims may be subject to clear time related limits, that can lead to greater apparent stability in some development periods
- Different treatments of tranches of claims, either due to claims management, legislative change or other source.
- Other sources of correlation amongst claims (economic conditions, claiming trends) might vary over time.

In our experience heteroscedasticity is a significant factor in many continuance rate and payment triangles. This means that plug-in estimates of outstanding claim uncertainty should be used with some caution. The issue motivates a couple of treatments when using GLMs:

- The weight of various cells can be altered so that less weight is placed on areas of higher variance. This will simultaneously fix the distribution estimation and increase the estimation accuracy.
- Less emphasis is placed on the traditional variance estimates and more reliance is placed on other tools such as the bootstrap, which should better recognise the true variability seen in the dataset.

4.5 Link functions Most GLMs that we have seen use log links for both the continuance and payment models. This approach has some intuitive appeal. First, a log link is canonical for a Poisson distribution. Second, the log link leads to a multiplicative model, which is intuitively attractive for many payment models. Ultimately the choice of link is of small but significant consequence - choosing a incorrect link can lead to slightly less accurate estimates and require extra interactions between variables. However in practice the mean estimates produced by well fit models under alternative links are often comparable. In simple cases such as Models 1 and 2 above, the mean estimates will generally be independent of the choice of link.

However, there are some reasons why exploring alternative link functions for the continuance rate model might justified. In more complex models where factors such as age are allowed for, there may be more interactions and parameters which will make that the choice of link function is more important. The log link permits estimates significantly larger than 1, since the multiplicative structure can lead to particularly high estimates in some cells. If the rate of reactivations is zero, than continuance rates are bounded above by 1, so enforcing this via the adoption of a logit or probit link makes sense. However in practice, the existence of reactivations mean that continuance rates above 1 are observed, which can lead to convergence issues under these links.

Two links that help limit the occurrence of large continuance rate estimates are:

- A scaled logit function $g(x) = \frac{1}{k}log\{(x)/(k-x)\}$, where the scaling factor k is the maximum permitted continuance rate throughout
- A tempered log link, that grows at some rate slower than exponential once the mean is greater than 1. For example:

$$g^{-1}(\eta) = \begin{cases} exp(\eta) & \text{if } \eta < 0 \\ \eta + 1 & \text{otherwise} \end{cases}$$

The difference between this, a scaled logit link, and a standard log link is shown in Figure 1.

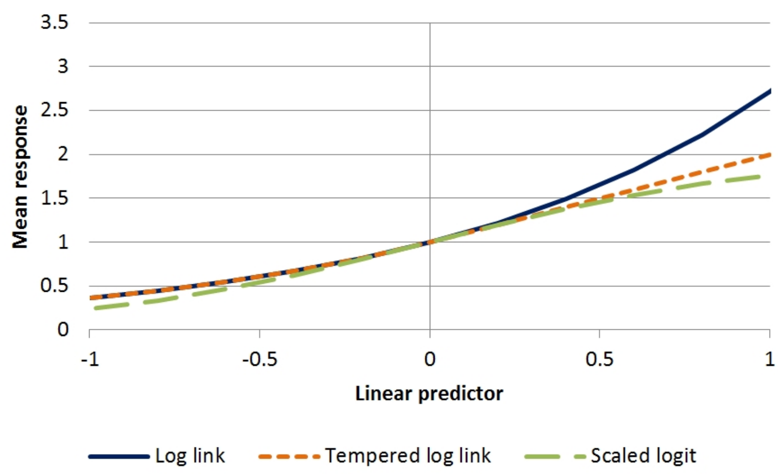

Figure 1: Standard log link, tempered log link and scaled logit link functions

4.6 Incorporating continuance rate information into the payment model assumptions Suppose that the payment stream represented primarily weekly ongoing benefits to each claimant (as opposed to lump sums), and that the rate at which claims deactivate within each time period is roughly uniform. Then the results of the continuance rate estimation should hold information relevant to the payments model. If $\mathcal{P}_{ij}$ is the average amount paid to a claimant who continues to be active into the next development period, then P_{ij} would be expected to be equal to $\mathcal{P}_{ij}$ for those that continue, and $\mathcal{P}_{ij}/2$ for those who deactivate. So:

$$P_{ij} \approx \mathcal{P}_{ij} \left\{ c_{ij} + \frac{1}{2}(1 - c_{ij}) \right\} = \mathcal{P}_{ij} \left(\frac{1}{2} + \frac{c_{ij}}{2} \right)$$

The observed factors $\log(\frac{1}{2}+c_{ij}/2)$ can be used as offsets to the payments model, resulting in easier estimation of the P_{ij}, since the variability due to discontinuance has been removed.

A similar approach can be used in traditional (non-GLM) contexts by scaling average payments

$$P'_{ij} = P_{ij} \left(\frac{1}{2} + c_{ij}2 \right)$$

before fitting the triangle. The adjustment helps to standardise the payment data; it corrects for unusually high or low deactivations before modelling.

4.7 Enforcing plausible continuance rates One area where GLM based approaches still lag behind traditional approaches is the ability to easily modify selected continuance rate (or payment level) assumptions. This typically arises when the analyst judges there are factors affecting future continuance rates not reflected in the historical data. While ad hoc judgements are usually less desirable than historical evidence, there are circumstances where they are unavoidable. For example, continuance rates can be affected by changes in legistlation or claims management practices.

We propose two options to address the issue. First, the GLM can be implemented in a spreadsheet, so that a subsequent ad hoc selection can be easily applied. This implementation can be genuine (software such as Microsoft Excel has robust solver routines that allow maximum likelihood optimizations to be performed), or via GLM output from statistical software directly linking into a spreadsheet.

Second, there are ways to factor in judgements into the GLM itself. One of the more direct means is to “penalise” the likelihood function from departing significantly from prior beliefs that the analyst may impose. This can be interpreted in a Bayesian or mixed model sense (see for example McCulloch and Neuhaus, 2001 or West et al, 2007), or a frequentist approach to variance reduction (see the penalised regression literature such as Tibshirani 1996, or Chapters 3 and 4 or Hastie et al 2009). Suppose that $L(N, \tilde{c})$ is the likelihood function that depends on (among other things) the observed numbers $N = \{n_{ij}\}$ and the vector of continuance rates for the latest diagonals $\tilde{c} = \{c_1, \ldots c_J\}$. Suppose further that $\tilde{c}^* = \{c_1^*, \ldots, c_J^*\}$ is a set of pre-determined continuance rate assumptions that the analyst prefers. Then minimising the penalised expression:

$$-L(N, \tilde{c}) + \sum_{j=1}^{J} \lambda_j (c_j - c_j^*)^2$$

with penalty factors $\lambda_j \geq 0$ will tend to push continuance factors towards the $\tilde{c}^*$. This approach is attractive because the integrated penalisation allows the statistical advantages of the GLM to be retained while giving some flexibility as the the adopted factors. The λ_j can be varied to influence how heavily selections should be weighted towards the pre-judged estimates - as the λ_j get very large, the fitted c_j will approach the c_j^*. However there are challenges to this approach; statistical software has been slow to recognise the potential of penalised regression, and good starting choices of λ_j are not always available.

A similar approach can be used for payment models to impose competing beliefs on adopted projection levels.

5 Concluding comments

We believe that there are significant advantages to using a GLM approach to reserving problems such as PPAC models. While there is a cost in terms of complexity and ad hoc flexibility, these are increasingly minor as we discover the correct ways to setup models, and as the supporting software evolves. This paper addresses a host of the pactical issues that allow an analyst to avoid a few of the pitfalls of GLMs, further enabling their adoption.

Acknowledgements

This research has been supported by a grant from the Institute of Actuaries Australia, which I acknowledge with thanks.

References

Clark, D.R., & Thayer, C.A. (2004). A primer on the exponential family of distributions. *Casualty Actuarial Society Spring Forum* 117–148.

Dobson, A.J. (2002). *An introduction to generalized linear models.* CRC Press LLC.

England, P.D., & Verrall, R.J. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. *Insurance: mathematics and economics*, **25(3)**, 281-293

England, P.D., & Verrall, R.J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, **8(03)**, 443–518.

Fahrmeier, L. & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models.* Springer.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). sl The elements of statistical learning (Vol. 2, No. 1). New York: Springer.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *Astin bulletin*, **23(02)**, 213–225.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

McCulloch, C.E. & Neuhaus, J.M. (2001). *Generalized linear mixed models.* John Wiley & Sons, Ltd.

Taylor, G.C. (1986). *Claims reserving in non-life insurance* (Vol. 1). Elsevier Science Ltd.

Taylor, G.C. (2000). *Loss reserving: an actuarial perspective.* Boston: Kluwer Academic.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

West, B.T., Welch, K.B., and Galecki, A.T. (2006). *Linear mixed models: a practical guide using statistical software.* CRC Press.

6 Appendix: Technical details

6.1 Theoretical results This section contains theoretical results supporting the arguments and claims made in the main body of the paper.

We adopt the same formulation for GLMs as Chapter 6 of Taylor (2000). In particular, the response Y_{ij} (either the continuance c_{ij} or payment level P_{ij}) has the distribution arising from an exponential family:

$$p(y) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{a_{ij}(\phi)} + c(y_{ij}, \phi)\right\} \tag{6.1}$$

with $a_{ij}(\phi) = \phi/w_{ij}$ and the mean equal to $b'(\theta_{ij}) = g^{-1}(X^T\beta)$.

We define a GLM as *disjoint* if every predictor vector X_{ij} related to cell (i, j) comprises of a single 1 and otherwise zeros. Models 1, 2 and 3 in the paper all fit this definition. The definition could be broadened to those that can be reformulated as disjoint using a series of variable transformations and linear combinations. The first result is a relatively straightforward property of GLMs, included for completeness.

Result 1. *If a GLM formulation is disjoint, then every parameter estimate is the weighted mean of those observations for which the corresponding variable is nonzero.*

Proof. The disjoint formulation means that the maximum likelihood estimation can be treated as a series of independent estimations on each of the groups of observations corresponding to each variable. That is, if $\mathcal{H}_m$ is the set of observations where the mth term of X_{ij} is 1, then the parameter corresponding to X_{ij} can be viewed as a maximum likelihood estimation of a intercept only model on $\mathcal{H}_m$. Thus it is sufficient to prove that for an intercept model (where $\theta_{ij} = \theta$ a constant), the estimate recovered is the weighted mean.

Using (6.1), we see that the log-likelihood is

$$L = \sum_{i,j\in\mathcal{H}} \phi^{-1} w_{ij}\{y_{ij}\theta_{ij} - b(\theta_{ij})\} + c(y_{ij}, \phi)$$

Setting $\theta_{ij} = \theta$, differentiating by θ and setting to zero to maximise the log-likelihood,

$$0 = \sum_{i,j\in\mathcal{H}} \phi^{-1} w_{ij}\{y_{ij} - b'(\theta_{ij})\} \tag{6.2}$$

and so

$$g^{-1}(X^T\beta) = b'(\theta_{ij}) = \frac{\sum w_{ij}y_{ij}}{\sum w_{ij}}.$$

If $g^{-1}(X^T\beta)$ is the intercept model so the GLM estimate is a constant $g^{-1}(X^T\beta) = g^{-1}(\beta)$, then this shows it must be the overall weighted average as required. □

Result 1 has useful implications for the paper - in particular it shows that the simple formulations of Models 1, 2 and 3 lead to parameters that recover the volume weighted averages as claimed. This establishes the equivalence to traditional PPAC model fits.

Result 2 demonstrates the increased volatility of running a deactivation/reactivation model compared to a single net deactivation model. But first we define:

- o_{ij} is the number of inactive claims in cell (i, j).
- For the deactivation/reactivation model, let d_j be the (gross) continuance rate (applied to $n_{i,j-1}$) and r_j the reactivation rate (applied to $o_{i,j-1}$).

We use capitalised versions (such as N_{ij} for n_{ij}) to denote random variables for cell (i, j). We assume a standard chain ladder setup where c_j if fixed for each development year and estimates of c_j are the volume weighted averages over $\mathcal{H}$.

Result 2. *Suppose that $d_j > 0.5$ and that standard chain ladder estimates apply and are made for c_j, d_j and r_j. Then the variance of one development period projection* $\mathrm{var}(N_{ij}|N_{i,j-1}, O_{i,j-1})$ *is lower for the net continuance rate model, compared to the deactivation/reactivation model.*

Proof. For the net deactivation model the number of actives in cell (i, j), N_{ij}, depends on $n_{i,j-1}$ and the estimate of c_j, which can be regarded as a binomial variable with $n_{i,j-1}$ counts and mean c_j which is estimated on $K := \sum_{i:i,j\in\mathcal{H}} n_{i,j-1}$ observations.

$$\begin{aligned}\mathrm{var}(N_{ij}|N_{i,j-1}) &= \mathrm{var}(n_{i,j-1}C_j) \\ &= n_{i,j-1}^2 \frac{c_j(1-c_j)}{K}\end{aligned}$$

Now the net continuance rate is actually a combination of the deactivation and reactivation rates,

$$c_j = d_j + \frac{o_{i,j-1}}{n_{i,j-1}} r_j \,.$$

Thus

$$\begin{aligned}\mathrm{var}(N_{ij}|N_{i,j-1}) &= n_{i,j-1}^2 \frac{(d_j + \frac{o_{i,j-1}}{n_{i,j-1}} r_j)(1 - d_j - \frac{o_{i,j-1}}{n_{i,j-1}} r_j)}{K} \\ &= \frac{n_{i,j-1}^2 d_j(1-d_j)}{K} + \frac{n_{i,j-1} o_{i,j-1} r_j (1-2d_j)}{K} - \frac{o_{i,j-1}^2 r_j^2}{K} \\ &< \frac{n_{i,j-1}^2 d_j(1-d_j)}{K}\,, \end{aligned} \tag{6.3}$$

where the last uses the $d_j > 0.5$ condition.

For the deactivation/reactivation model, if we similarly define $L := \sum_{i:i,j\in\mathcal{H}} o_{i,j-1}$,

$$\begin{aligned}\mathrm{var}(N_{ij}|N_{i,j-1}, O_{i,j-1}) &= \mathrm{var}(n_{i,j-1}D_j + o_{i,j-1}R_j) \\ &= n_{i,j-1}^2 \frac{d_j(1-d_j)}{K} + o_{i,j-1}^2 \frac{r_j(1-r_j)}{L} \end{aligned} \tag{6.4}$$

Comparing (6.3) and (6.4) gives the desired inequality. □

Three comments on Result 2:

- The result proves the extra volatility over a single projection period, the projection process of the chain ladder will tend to compound this extra variance over the full completion of the triangle. However the situation is more complex due to the extra information in the triangle of inactive claims, and so the simple inequality above is not necessarily true over multiple periods.

- The inequality is fairly "loose", in that the differences will be significant if the discarded terms are large. In particular, as the number of reactivations (the product of o_{ij} and r_j) grow, the extra volatility can be significant and the $d_j > 0.5$ condition unnecessary.

- The $d_j > 0.5$ condition is a fairly weak assumption in practice, given that PPAC models tend to be longer tail projections with continuance rates closer to one.

We further illustrate the implications of Result 2 with a simulated example in Section 6.2 below.

The third result establishes the legitimacy of using a small non-zero entry in the zero cells of an actives triangle. This idea was introduced in Section 3.2. Let $\mathcal{H}^* \subset \mathcal{H}$ be the cells (i, j) of the triangle such that $n_{i,j-1} = 0$.

Result 3. *Set $n_{i,j-1} = \delta > 0$ for all (i, j) in $\mathcal{H}$. Then as $\delta \to 0$, the log-likelihood and resulting GLM parameters converge. Further, if the GLM is disjoint then the estimates converge to the volume weighted average continuance in each region.*

Proof. As with Result 1, the disjoint condition means that it is sufficient to prove that the intercept model converges to the weighted average continuance rate. Using (6.2) and the setup for response and weight in Section 2.4, the derivative of the log-likelihood expression becomes

$$\begin{aligned} 0 &= \sum_{i,j \notin \mathcal{H}^*} \phi^{-1} n_{i,j-1} \{c_{ij} - b'(\theta_{ij})\} + \sum_{i,j \in \mathcal{H}^*} \phi^{-1} \delta \{n_{ij}/\delta - b'(\theta_{ij})\} \\ &= \phi^{-1} \sum_{i,j \in \mathcal{H}} \{n_{i,j} - n_{i,j-1} b'(\theta_{ij})\} - \delta \phi^{-1} \sum_{i,j \in \mathcal{H}^*} b'(\theta_{ij}) \end{aligned}$$

As $\delta \to 0$ the first sum remains constant while the second tends to zero, as required. For the intercept model, taking the limit and solving for $g^{-1}(\beta) = b'(\theta)$ gives:

$$g^{-1}(\beta) = b'(\theta) = \frac{\sum_{i,j \in \mathcal{H}} n_{ij}}{\sum_{i,j \in \mathcal{H}} n_{i,j-1}},$$

recovering the volume weighted average continuance rate as claimed. □

6.2 Simulation results To further illustrate Result 2, we have constructed a simple chain ladder problem for an active claim triangle:

- The triangle has 10 development years, plus a year 0.
- At year 0 in each row there are 100 claims active and 100 claims inactive
- (Gross) Continuance rates of 85% and reactivation rates of 1% are assumed in all development years.

We calculate the standard chain ladder estimates and complete the triangle, using both a net continuance rate model and a deactivation-reactivation model, as discussed in Section 3.1. Our metric of interest is the standard deviation of the sum of the final (eleventh) row of the triangle, $\sum_{j=1}^{10} n_{11,j}$. This will be roughly proportional to the liability for the most recent accident year, so is a good proxy for understanding modelling volatility.

Over 5,000 simulations the observed standard deviation in the net continuance rate model was 16.3, while for the deactivation-reactivation model it was 17.5, about 7% higher. The distribution of spreads is shown in Figure 2. The fatter tails of deactivation-reactivation model can be seen, also it also shows less systematic bias (under-estimation of the mean). This demonstrates that there is a significant amount of extra volatility introduced by the dual model structure.

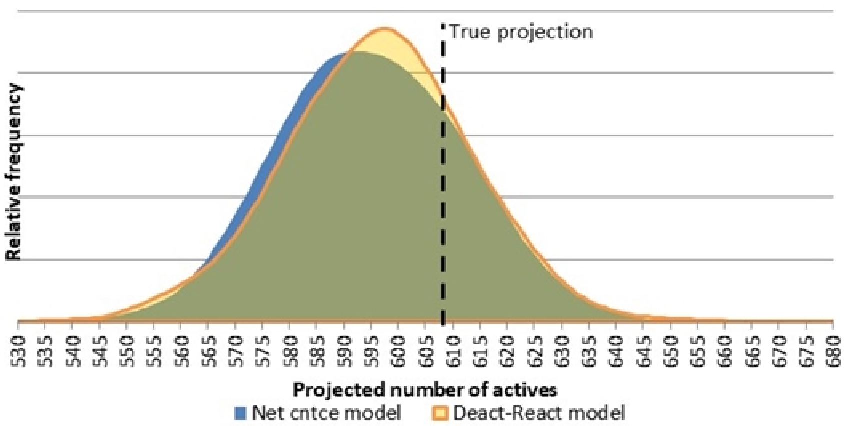

Figure 2: Distribution of chain ladder estimates for net continuance and deactivation-reactivation models

We can see how this ratio between standard deviations varies as simulation parameters are altered. Figure 3 shows the results if we change the gross continuance rate, the reactivation rate and the number of starting inactive claims. In each case we can see the standard deviation of the deactivation-reactivation model is around double, with some trends visible:

- The gap in standard deviation between the two models is smaller when the gross continuance rate is lower. This is some way accords with the theoretical result, and is probably due to the deactivation model dominating the triangle, leaving less scope for the reactivations to increase variability.

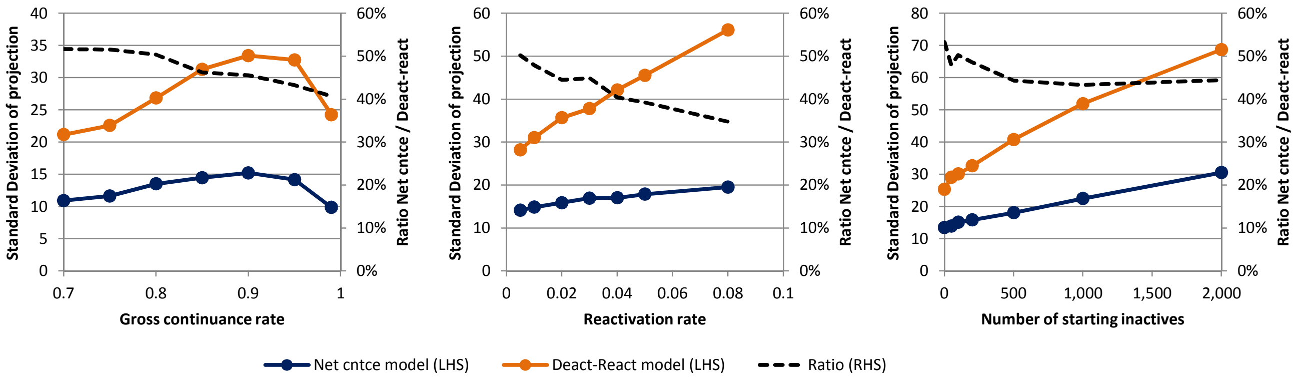

Figure 3: Results for simulation as parameters are varied

- The gap is smaller as the reactivation rate shrinks. If the reactivation rate was zero, then there will be no difference at all, as the models would be equivalent
- The gap was largest when the number of starting actives is close to the starting number of inactives. In particular, the variance of the net continuance rate model is higher when there are zero starting inactive claims.

In practice these types of comparisons are further complicated by the other features of the triangle and model selection. However, the general pattern of higher variance with the deactivation-reactivation model appears a plausible rule of thumb.